TITLE OF THE INVENTION

GRAPHEME-PHONEME CONVERSION

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0001] The invention relates to a method, a computer program product, a data medium and a computer system for grapheme-phoneme conversion of a word which is not contained as a whole in a pronunciation lexicon.

2. Description of the Related Art

[0002] Speech processing methods in general are known, for example, from US 6 029 135, US 5 732 388, DE 19636739 C1 and DE 19719381 C1. In a speech synthesis system, the script-to-speech conversion or grapheme-phoneme conversion of the words to be spoken is of decisive importance. Errors in sounds, syllable boundaries and word stress are directly audible, can lead to incomprehensibility and can, in the worst case, even distort the sense of a statement.

[0003] The best quality speech recognition is obtained when the word to be spoken is contained in a pronunciation lexicon. However, the use of such lexica causes problems. On the one hand, the number of entries increases the search outlay. On the other hand, it is precisely in the case of languages such as German that it is impossible to cover all words in a lexicon, since the possibilities of forming compound words are virtually unlimited.

[0004] A morphological decomposition can provide a remedy in this case. A word which is not found in the lexicon is decomposed into its morphological constituents such as prefixes, stems and suffixes and these constituents are searched for in the lexicon. However, a morphological decomposition is problematical precisely in the case of long words, because the number of possible decompositions rises with the word length. However, it requires an excellent knowledge of the word formation grammar of a language. Consequently, words which are not found in a pronunciation lexicon are transcribed with out-of-vocabulary methods (OOV methods), for example, with the aid of neural networks. Such OOV treatments are, however, relatively compute-intensive and generally lead to poorer results than the phonetic conversion of whole words with the aid of a pronunciation lexicon. In order to determine the pronunciation of a word which is not contained in a pronunciation lexicon, the word can also be decomposed into

subwords. The subwords can be transcribed with the aid of a pronunciation lexicon or an OOV method. The partial transcriptions found can be appended to one another. However, this leads to errors at the break points between the partial transcriptions.

SUMMARY OF THE INVENTION

[0005] It is an object of the invention to improve the joining together of partial transcriptions. This object is achieved by a method, a computer program product, a data medium and a computer system in accordance with the independent claims.

[0006] In this case, a computer program product is understood as a computer program as a commercial product in whatever form, for example on paper, on a computer-readable data medium, distributed over a network, etc.

[0007] According to the invention, in the grapheme-phoneme conversion of a word which is not contained as a whole in a pronunciation lexicon, the first step is to decompose the word into subwords. A grapheme-phoneme conversion of the subwords is subsequently carried out.

[0008] The transcriptions of the subwords are sequenced, at least one interface being produced between the transcriptions of the subwords. Phonemes, bordering on the interface, of the subwords are determined.

[0009] It is possible in this case to take account only of the last phoneme of the subword situated upstream of the interface in the temporal sequence of the pronunciation. However, it is better when both this phoneme and the first phoneme of the following syllable are selected for the special treatment according to the invention. Even better results are achieved when further bordering phonemes are included, for example, one or two phonemes upstream of the interface and two downstream of the interface.

[0010] Subsequently, those graphemes of the subwords are determined which generate the phonemes bordering on the at least one interface. This can be performed by using a lexicon which specifies which graphemes generated these phonemes. How the lexicon is to be created is set forth in Horst-Udo Hain: "Automation of the Training Procedures for Neural Networks Performing Multilingual Grapheme to Phoneme Conversion", Eurospeech 1999, pages 2087-2090.

[0011] Hereafter, the grapheme-phoneme conversion of the specific graphemes is recalculated in the context, that is to say, as a function of the context, of the respective interface. This is possible only because it is clear which phoneme has been created by which grapheme or graphemes.

[0012] The interfaces between the partial transcriptions are therefore treated separately. If appropriate, changes to the previously determined partial transcriptions are undertaken. An advantage of the invention which is not inconsiderable for a speech synthesis system is the acceleration of the calculation. Whereas neural networks require approximately 80 minutes for converting the 310 000 words of a typical lexicon for the German language, this is performed in only 25 minutes with the aid of the approach according to the invention.

[0013] In an advantageous development of the invention the grapheme-phoneme conversion of the graphemes can be recalculated in the context of the respective interface by using a neural network. A pronunciation lexicon has the advantage of supplying the "correct" transcription. It fails, however, when unknown words occur. Neural networks can, by contrast, supply a transcription for any desired character string, but make substantial errors in this case, in some circumstances. The development of the invention combines the reliability of the lexicon with the flexibility of the neural networks.

[0014] The transcription of the subwords can be performed in various ways, for example by using an out-of-vocabulary treatment (OOV treatment). A very reliable way consists in searching for subwords for the word in a database which contains phonetic transcriptions of words. The phonetic transcription recorded in the database for a subword found in the database is then selected as transcription. This leads to useful results for most words or subwords.

[0015] If, in addition to the subword found, the word has at least one further constituent which is not recorded in the database, this constituent can be phonetically transcribed by using an OOV treatment. The OOV treatment can be performed by a statistical method, for example by a neural network or in a rule-based fashion, e.g., using an expert system.

[0016] The word is advantageously decomposed into subwords of a certain minimum length, so that subwords as large as possible are found and correspondingly few corrections arise.

[0017] The invention is explained in more detail below with the aid of exemplary embodiments which are illustrated schematically in the figures.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a computer system suitable for grapheme-phoneme conversion; and

Figure 2 shows a schematic of the method according to the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0018] Figure 1 shows a computer system suitable for grapheme-phoneme conversion of a word. The system has a processor (CPU) 20, a main memory (RAM) 21, a program memory (ROM) 22, a hard disk controller (HDC) 23, which controls a hard disk 30 and an interface (I/O) controller 24. The processor 20, main memory 21, program memory 22, hard disk controller 23 and interface controller 24 are coupled to one another via a bus, the CPU bus 25, for the purpose of exchanging data and instructions. Furthermore, the computer has an input/output (I/O) bus 26 which couples the various input and output devices to the interface controller 24. The input and output devices include, for example, a general input and output (I/O) interface 27, a display 28, a keyboard 29 and a mouse 31.

[0019] Taking the German word "uberflüssigerweise" as an example for grapheme-phoneme conversion, the first step is to attempt to decompose the word into subwords which are constituents of a pronunciation lexicon. A minimum length is prescribed for the constituents being sought in order to restrict the number of possible decompositions to a sensible measure. Six letters have proved to be sensible in practice as minimum length for the German language.

[0020] All the constituents found are stored in a chained list. In the event of a plurality of possibilities, use is always made of the longest constituent or the path with the longest constituents.

[0021] If not all parts of the word are found as subwords in the pronunciation lexicon, the remaining gaps in the preferred exemplary embodiment are closed by a neural network. By contrast with the standard application of the neural network, in case of which the transcription must be created for the entire word, the task in filling up the gaps is simpler because at least the left-hand phoneme context can be assumed as certain since it does originate, after all, from the pronunciation lexicon. The input of the preceding phonemes therefore stabilizes the output of

the neural network for the gap to be filled, since the phoneme to be generated depends not only on the letters, but also on the preceding phoneme.

[0022] A problem in mutually appending the transcriptions from the lexicon and in determining the transcription for the gaps by a neural network consists in that in some cases the last sound of the preceding, left-hand transcription has to be changed. This is the case with the considered word "überflüssigerweise". It is not found in the lexicon as a whole, but the subword "überflüssig" and the subword "erweise" are.

[0023] For the purpose of better distinction, graphemes are enclosed below in pointed brackets <>, and phonemes in square brackets [].

[0024] The ending <-ig> at the end of a syllable is spoken as [IC], represented in the SAMPA phonetic transcription, that is to say as [I] (lenis short unrounded front vowel) followed by the "Ich" sound [C] (voiceless palatal fricative). The prefix <er-> is spoken as [Er], with an [E] (lenis short unrounded half-open front vowel, open "e") and an [r] (central sonorant).

[0025] In the case of simple chaining of the transcriptions, it is sensible to insert automatically between the two words a syllable boundary represented by a hyphen "-". The result as overall transcription of the word <über-flüssigerweise> is therefore:

[0026] [y: - b6 - flY - sIC - Er - val - z@]

[0027] instead of, correctly,

[0028] [y: - b6 - flY - sI - g6 - val - z@]

[0029] with a [g] (voiced velar plosiv) and a [6] (unstressed central half-open vowel with velar coloration) as well as a displaced syllable boundary. This would mean that sound and syllable boundary were wrong at the word boundary.

[0030] A remedy may be provided here by using a neural network to calculate the last sound of the left-hand transcription. In this case, however, the question arises as to which letters at the end of the left-hand transcription are to be used to determine the last sound.

[0031] A special pronunciation lexicon is used for this decision. The special feature of this lexicon consists in that it contains the information as to which grapheme group belongs to which sound. How the lexicon is to be created is set forth in Horst-Udo Hain: "Automation of the

Training Procedures for Neural Networks Performing Multilingual Grapheme to Phoneme Conversion", Eurospeech 1999, pages 2087-2090.

[0032] The entry for "überflüssig" has the following form in this lexicon:

ü	-	b	er	-	f	l	ü	-	ss	i	g
y:	-	b	6	-	f	l	y	-	s	l	C

[0033] It is therefore possible to determine uniquely from which grapheme group the last sound has arisen, specifically from the <g>.

[0034] The neural network can now use the right-hand context <erweise> now present to make a new decision on the phoneme and syllable boundary at the end of the word. The result in this case is the phoneme [g], in front of which a syllable boundary is set.

[0035] The syllable boundary is now at the correct position and the <g> is also transcribed as [g] and not as [C].

[0036] The first sound of the right-hand transcription is redetermined using the same scheme. The correct transcription for <er-> of <erweise> is at this point [6] and not [Er]. Here, two sounds precisely are to be checked, for which reason two sounds are always checked in the preferred exemplary embodiment.

[0037] The correct phonetic transcription at this interface is obtained as a result.

[0038] Further improvements are to be achieved when use is made for the purpose of filling up the transcription gaps, not of the standard network, which has been trained to convert whole words, but of a network specifically trained to fill up the gaps. At least in the cases in which the right-hand phoneme context is also present, a specific network is on offer which uses the right-hand phoneme context to decide on the sound to be generated.